预印本。正在审阅。1–17

arXiv:2302.02011v1 [cs.RO] 3 Feb 2023

从被动人类视频进行零样本机器人操作

Homanga Bharadhwaj[1,2], **Abhinav Gupta**[1], **Shubham Tulsiani**[1,*], **Vikash Kumar**[2,*]
[1] 机器人研究所、卡内基梅隆大学 [2]*Meta AI*

[*]共同贡献

摘要

我们能否仅通过观看人类在不同非结构化环境中执行任意任务的视频，学会完成日常任务的机器人操作？不同于广泛采用的学习特定任务行为或直接模仿人类视频的策略，我们开发了一个框架，用于从人类视频中提取与智能体无关的动作表征，并在部署时将其映射到智能体的具身形态。我们的框架基于给定场景初始图像，预测合理的人手轨迹。在互联网上收集的多样化人类视频上训练该预测模型后，我们对模型进行适当转换，使其适应机器人的具身形态，随后将训练好的模型零样本地部署于现实世界的机器人操作任务中。这一简单策略使我们能够完成打开和关闭抽屉、推物以及工具使用等粗略操作任务，而无需获取任何领域内的机器人操作轨迹。我们的现实世界部署结果为从多样化的任意人类活动视频中获取动作预测信息，并将其用于未见场景中的零样本机器人操作，确立了一个强有力的基线。 [1]

关键词： 从人类视频中学习，机器人操作

1. 引言

我们人类在日常生活中能够毫不费力地完成大量操作任务，例如打开橱柜、切菜、倒咖啡、转动旋钮等。在快速发展的（数据驱动）机器人学习领域，一个常见目标是开发能够在非结构化环境中执行多种任务的智能体。基于深度强化学习的方法Kalashnikov 等人(2021年) 提供了一种框架，使机器人能够通过优化相应的奖励，不断提升执行通用任务的能力。然而，样本（非）效率低下、需要在线交互，以及奖励设计和环境重置难度较大，通常使这类方法的应用局限于结构化环境中的特定任务。另一种方法是直接利用专家提供的机器人演示来学习动作策略Pomerleau (1989年)；Finn 等人(2017年)。原则上，在不同环境和任务中收集的演示能够帮助学习所需的多样化行为，但这类演示通常通过远程操作或动觉教学收集，因此难以大规模扩展。此外，这些方法所需的在线交互和机器人演示都预先要求定义任务，并且受限于变化很少的实验室环境；因此，它们与训练能够在未知非结构化环境中有效工作的通用机器人所需的多样化数据相去甚远。

是否存在另一种数据来源，能够支持机器人操作的学习？在这项工作中，我们展示了，人类在完成日常任务时与物体交互的视频，可以作为一种

1. hbharadh@cs.cmu.edu视频 https://sites.google.com/view/human-0shot-robot

© H.B.A.G.S.T.K. .

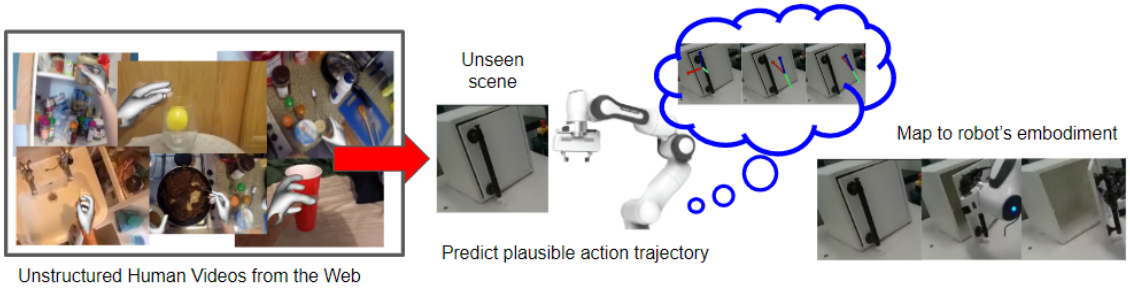

图 1：从被动人类视频进行零样本机器人操作的所提框架概述。我们的方法基于根据网络上的人类视频中的场景图像预测手部轨迹，并将预测结果转换为机器人在实验室中面对未见场景时进行零样本机器人操作所需的信息。

随手可得的大规模数据来源。具体而言，我们展示了，对人类-物体交互视频中人手运动的建模，能够零样本地预测机器人为实现其环境中的目标应采取的动作。我们提出了一个框架：给定一张初始图像，模型学习预测后续帧中人手在场景内动作的运动序列。具体来说，模型预测的是从初始图像开始、未来 2 秒时间范围内人手手掌中心某一点的六维位姿。由于场景中可能存在多种不同且合理的动作，我们将模型设计为随机模型，并使用现有数据集（如 Epic-Kitchens）中的多样化第一人称视角视频对其进行训练Damen 等人(2018年）。得益于人类视频的规模和多样性，该预测模型使我们能够在包含新物体的不同物理环境中进行零样本动作预测。我们将这类人类视频称为被动的，因为它们并非为了机器人学习而主动采集，也没有按任务级别进行分类；我们的框架则被称为零样本，因为没有实验室内演示或微调。图1 展示了我们从被动人类视频进行零样本机器人操作的框架概览

我们提出了该框架的两个版本：无条件版本和目标条件版本。无条件模型根据场景预测可能的轨迹，对应于实现该场景中物体的不同可供性，这有助于后续探索。目标条件模型则进一步以目标图像作为提示，预测能够到达指定目标的轨迹，有助于实现特定目标行为。为了根据预测结果操控机器人，我们将其转换到机器人的末端执行器空间，并使用简单的逆运动学控制器来执行轨迹运动。我们展示了这种简单方法的可靠性：在无条件操控烤面包机、抽屉、碗等物体时，成功率约为 40%–60%（见图 3）；在目标条件操控中，成功率约为 30%–40%。最后，我们指出，结合某种形式的特定环境数据微调，必定能提升精细操作任务的表现；不过，当前框架为仅凭与机器人无关的人类数据实现零样本操作提供了一个出乎意料的强大基线。

2. 相关工作

我们探讨了以往通过检测手部姿态、语义任务、视觉特征提取和活动预测来理解视频中人类活动的研究。随后，我们讨论了从人类和机器人执行不同活动的视频中学习机器人操作技能的论文。

扩展人-物交互理解。随着大规模数据集的发展，理解人类活动近来备受关注，Goyal 等人(2017年）；Das 等人(2013年）；De la Torre 等人(2009年）；Grauman 等人(2022年）这些数据集记录了人类在厨房进行烹饪相关活动的视频Damen 等人(2018年）；Li 等人(2018年），还包括涉及物体操作的短片段Goyal 等人(2017年），以及记录家庭内外活动的更多样、更长的视频Grauman 等人(2022年）；Shan 等人(2020年）。

基于这些大规模人类视频数据集，多项先前研究致力于理解人-物交互 Shan 等人（2020）；Ye 等人（2022）。具体而言，先前研究探讨了物体姿态估计 Kehl 等人（2017）；Rad 和 Lepetit（2017）；Xiang 等人（2018）；Hu 等人（2019）；He 等人（2020）、手部姿态估计 Zimmermann 和 Brox（2017）；Iqbal 等人（2018）；Spurr 等人（2018）；Ge 等人（2019）；Baek 等人（2019）；Boukhayma 等人（2019）；Hasson 等人（2019）；Kulon 等人（2020）；Liu 等人（2021）、全身姿态估计 Rong 等（2020），以及手-物联合姿态估计 Ye 等人（2022）；Garcia-Hernando 等人（2018）；Hampali 等人（2020）；Liu 等人（2021）；Chao 等人（2021）。这些研究也得到了真实物体三维扫描数据集的有力补充，例如人类通常会与之交互的 YCB 数据集 Calli 等人（2015）和 Google Obbject Scans 数据集 Downs 等人（2022）。另一类研究则探讨了从视频中学习交互热点 Nagarajan 等人（2019）；Liu 等人（2022）；Goyal 等人（2022），并预测合理抓取姿态 Mo 等人（2021）；Brahmbhatt 等人（2019）。

在这些主要聚焦于视觉理解的进展基础上，我们的工作着重于通过利用大规模人类视频被动采集数据集来闭合视觉-机器人闭环。我们学习合理的手部轨迹以与物体交互，并将其应用于真实的机器人操作。在下一小节中，我们将概述我们的框架与从视频中学习操作技能的机器人学习方法有何不同。

从视频进行机器人操作 机器人学的近期发展广泛关注如何利用日益非结构化的数据。视觉模仿学习方法旨在从视觉观测和机器人动作数据集学习控制策略 Finn 等（2017）；Young 等（2020a）；Mandlekar 等（2018）。行为克隆 Pomerleau（1989）；Bain 和 Sammut（1995）；Ross 和 Bagnell（2010）；Bojarski 等（2016）；Torabi 等（2018a）与逆强化学习 Russell（1998）；Ng 等（2000）；Abbeel 和 Ng（2004）；Ho 和 Ermon（2016）；Fu 等（2017）；Aytar 等（2018）；Torabi 等（2018b）；Liu 等（2020）是这一领域中两种常见的方法，但由于通常需要高质量的域内专家机器人轨迹（由人类控制机器人采集），因此难以扩展到未见场景。为减少采集高质量机器人数据的需求，一些方法利用人类执行各种活动的视频 Schmeckpeper 等（2019）；Chang 等（2020）；Schmeckpeper 等（2020）；Shaw 等；Shao 等（2020）；Song 等（2020）；Young 等（2020b），通过模仿或强化学习来学习控制策略。然而，对于模仿学习而言，人类需要在不同场景中通过专用硬件接口采集数据，以进行视觉模仿 Young 等（2020b）；Song 等（2020），而这

难以扩展。基于强化学习的方法则需要仿真环境，以进行带有重置和奖励的在线交互，从而在学习过程中提供反馈Shao 等 (2020);Schmeckpeper 等 (2020). 与这些方法相比，我们无需任何专门采集的数据或用于学习的模拟器，因此能够零样本泛化到未知任务。

其他研究通过直接模仿场景中对应的人类视频，探索了机器人运动的学习方法 Peng 等（2018）；Pathak 等（2018）；Sharma 等（2018）；Sieb 等（2020）；Sivakumar 等（2022）；Garcia-Hernando 等（2020）；Xiong 等（2021）；Sermanet 等（2018）；Sharma 等（2019）；Smith 等（2019）；Peng 等（2020）；Bahl 等（2022）；Qin 等（2021）。其中一些方法 Smith 等（2019）；Xiong 等（2021）虽不需要大规模数据集，但要求机器人手臂和人类手臂的姿态近乎完美对齐。近期研究 Bahl 等（2022）通过从场景中图像修复去除人类，以构建奖励函数，缓解了对完美对齐的需求；但它们仍需针对每项任务通过强化学习进行在线微调，这在现实世界中成本高昂。DexMV Qin 等（2021）结合视觉与仿真流程，能够有效地将人类视频转化为灵巧手部运动，但训练需要多段实验室内采集的人类视频，且无法利用网络上已有的人类视频。

与这些先前的工作相比，我们从互联网上的人类视频中提取与智能体无关的表征，形式为手部轨迹。我们不再直接模仿视频流，而是学习一个以场景为条件的轨迹预测模型，该模型以单张图像观测作为输入，可直接用于零样本机器人操作。这使操作能力能够泛化到未见场景，而无需预先设定任务规范。

3. 方法

受人类对如何以有趣的方式操作场景具有直观理解这一点的启发，我们开发了一种计算框架，仅通过观察人类在各种非结构化环境中与物体互动的视频，就能赋予机器人类似的操作技能。具体而言，给定来自单一视点的场景初始图像 o_1，我们学习一个模型$p(a_{1:T}|o_1)$，用于预测人类在场景中互动时所采取的未来动作 $a_{1:T}$。我们专注于桌面操作场景，其中 $a_{1:T}$ 对应于人类右手可能出现的六维位姿。为了使该框架能够普遍适用于配备简单末端执行器（如两指夹爪）的机械臂，我们不对手部关节运动（即各关节相对于手腕的朝向）建模，而只预测手掌的位置和方向。在各种互联网视频上训练该预测模型后，我们将其应用于实验室环境，无需使用实验室数据进行额外微调，即可在固定摄像机拍摄的桌面场景中操作抽屉、橱柜和门等物体。

3.1. 学习预测未来手部轨迹

给定一个场景的图像，可能存在多种合理的轨迹来改变场景中的物体。为了捕捉这种多模态性，我们为 p（$a_{1:T}|o_1$）开发了一种随机模型，使得从模型中采样不同结果会产生不同的轨迹。我们不预测每个时间步 h_t的绝对手部姿态，而是预测增量动作 $a_t = h_t \ominus h_{t-1}$，其中 $\ominus$ 是一个合适的“差分算子”。这样可以减小模型预测的绝对幅度，从而降低预测误差。此外，为了确保时间相关性，该模型采用自回归方式，在预测下一个动作时，将轨迹中的先前动作作为输入。在

接下来的小节中，我们将介绍预测模型架构，以及用于训练模型的数据预处理。

3.1.1. 模型架构

手部预测模型是一种图像 Transformer Bertasius 等人 (2021年); Chen 等人(2020年), 采用编码器—解码器架构，并在末端设置一个以解码器特征输出为条件的条件变分自编码器手部轨迹预测头。我们在图 2中概述了该架构。与先前研究类似

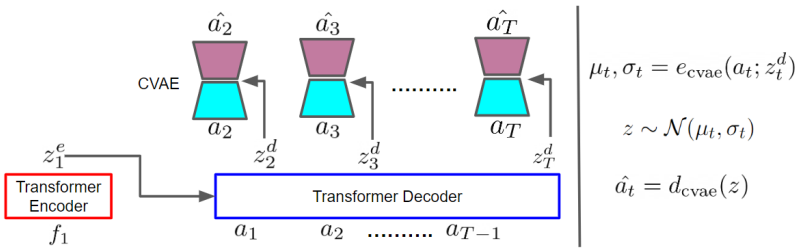

图 2： 手部轨迹预测模型的架构，展示了在给定初始场景特征时，无条件轨迹预测模型的结构。右侧展示了该模型的推理过程。详情请参见第 3.1.1

工作 Furnari 和 Farinella (2019年);Wang 等人 (2016年), 我们首先从初始图像中离线提取特征 o_1, 并将其记为 f_1。特征 f_1 由 Transformer 编码器编码 $\mathcal{E}$ 为潜在编码 z_1^e。Transformer 解码器 $\mathcal{D}$ 以潜在编码 z_1^e 和先前动作$a_{1:t-1}$ 为条件，输出时间步 t处用于动作预测的特征， z_t^d。

训练期间，条件变分自编码器以动作 a_t 和条件上下文 z_t^d作为输入，并输出重构结果 $\hat{a}_t$。推理期间，我们只需从该条件变分自编码器的先验分布中采样，将其与预测的上下文 z_t^d拼接，即可获得采样得到的增量动作 $\hat{a}_t$。获得增量动作后，我们可以将其还原为预测动作 $\hat{a}_t = \hat{a}_{t-1} \oplus \hat{a}_t \, \forall t > 1$，其中 $\oplus$ 是适当的“加法运算符”。对于目标条件模型，Transformer 编码器还会接收带有适当位置嵌入的目标图像嵌入 f_g，并输出 z_g^e。模型的具体细节见附录。

Transformer。 Transformer 编码器和解码器由多个堆叠的操作块组成，这些操作块包含注意力、多层感知机和层归一化。解码器块包含交叉注意力，其中查询与值令牌是编码后的代码 z_1^e。而编码器中的注意力块则全部为自注意力。

条件变分自编码器。 每个时间步的条件变分自编码器预测头由编码器 $e_{\text{cvae}}(\cdot)$和解码器$d_{\text{cvae}}(\cdot)$神经网络组成。以 Transformer 解码器预测的特征 z_t^d和当前增量动作 a_t 为输入条件，条件变分自编码器编码器输出高斯分布的均值 μ_t 和标准差 σ_t。条件变分自编码器解码器从该高斯分布中采样，并输出预测的增量动作 $\hat{a}_t$。形式化地， $\mu_t, \ \sigma_t = e_{\text{cvae}}(a_t; \ z_t^d)$ $z \sim \mathcal{N}(\mu_t, \ \sigma_t)$ $\hat{a}_t = d_{\text{cvae}}(z)$

训练损失。 整体训练损失根据条件变分自编码器在预测时域内所有时间步上逐步汇总的输出定义。 T

$$\mathcal{L} = \sum_{t=2}^{T} \left[||a_t - \hat{a}_t||^2 - D_{\text{KL}} \left(\mathcal{N}(\mu_t, \sigma_t) || \mathcal{N}(0, 1) \right) \right]$$

这一总体损失通过反向传播，传递至包含 Transformer 和条件变分自编码器（通过重参数化）的整个预测模型，且各阶段均没有中间监督。正如下一节所述，这使该方法普遍适用于真实场景下的人体视频，并减少训练数据的预处理开销。此外，



图 3：图中展示了我们在实验场景中摆放的几种物体配置，绿色箭头表示可能的运动；上排从左到右依次为：带垂直铰链的门、水果碗、放有蔬菜的砧板；下排依次为：带水平铰链的烤箱、上下叠放的两个抽屉，以及一个蔬菜滤篮。

所学习的模型与任务无关，因为人类视频片段中没有针对特定任务的区分；在未见场景中，它还能够以零样本方式执行合理的任务。

3.1.2. 训练数据生成

我们从 Epic-Kitchens 中选取 2 秒长的第一人称视角视频片段，这些片段记录人们进行日常家务活动，尤其是在厨房中的活动，例如烹饪、打开橱柜、将物体从一个位置移到另一个位置等。对于目标条件模型，片段的最后一帧图像作为输入模型的目标。为了获得预测时间窗口内各帧中手部的真实姿态，我们使用现成的 FrankMocap 模型进行手部跟踪（Rong 等（2020））。FrankMocap 输出弱透视相机参数（t_x, t_y, s）、所有手部关节的位置（x, y, z），以及手部相对于规范手的方向（α, β, γ）。我们仅考虑手掌中心，并记录其相对于 2 秒窗口起始时预测相机的六维位姿。总之，训练数据由成对的 $\{(o_1, h_{1:T})\}$ 构成，其中 o_1 是视频片段的第一帧图像，而 $h_t = (x_t, y_t, s_t, t_x, t_y, \alpha, \beta, \gamma)$ 是未来时间点 t 的手部姿态和相机参数。

3.2. 将轨迹映射到机器人的坐标系

使用多样化的互联网视频训练整体手部位姿预测模型 $p_\psi(a_{1:T}|o_1)$ 后，我们将其直接部署于实验室中的机器人操作任务。机器人通过固定摄像机观察场景图像，并可选择接收作为预测模型输入的目标图像。为了使用模型 $a_{1:T}$ 预测的动作来驱动机器人，我们需要将这些动作转换到机器人的世界坐标系，并将每个动作 a_t 视为末端执行器目标位姿，再使用低层控制器执行相应的运动。

场景中的相机经过了标定，因此内参矩阵 I 和外参矩阵 $[R, T]$ 是已知的。世界坐标系位于机器人基座处（机器人基座与桌面处于同一高度），且相机到桌面的高度已知并近似恒定。给定相机拍摄的场景 o_1，模型会预测增量动作 $a_{1:T}$，我们将其转换为绝对动作（详见第3.1.1节），然后通过逆投影变换将动作从相机坐标系转换到机器人的世界坐标系。我们的实验中，预测时域为 $T = 7$。得到动作序列的世界坐标 $\{(X_t,\ Y_t,\ Z_t,\ \alpha_t,\ \beta_t,\ \gamma_t)\}_{t=1}^{T}$ 后，我们使用逆运动学控制器执行相应的运动，使末端执行器在每个时间步到达所需的位置和方向。更多细节见附录。

4. 实验设计

通过实验，我们旨在探究以下研究问题：

- 对于未知任务，我们的无条件模型生成多样且合理的任务结果的能力如何？
- 对于已知任务，我们的目标条件模型和动作映射在解决任务方面的效果如何？

从定性角度看，我们展示了无条件模型在不同场景中执行的轨迹的多样性和合理性。对于定量评估，我们比较了机器人置于不同场景时无条件模型预测的成功率，以及在额外输入物体目标配置后目标条件模型的成功率。我们将其与一种三维场景流 Vedula 等人(1999);Teed 和 Deng (2021年) 基线进行比较，该基线使用 RAFT3D Teed 和 Deng (2021年) 进行预测

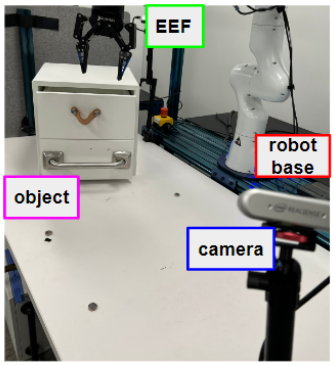

图 4：机器人工作区，展示了相机、场景中的一个物体、Franka 机械臂及其末端执行器。

初始图像与目标图像之间的场景流场，并利用主导流动方向引导机器人的运动。为测试在多样化数据上进行训练的重要性，我们将其与仅使用总训练数据 30% 进行训练的我们的方法版本进行比较。实验设置、物体、工具以及实验比较的更多细节见附录[2]。

实验设置：图 4 所示的工作空间是一张白色桌子，桌角之一放置着相机，机器人基座位于其相对的一侧。在图 3中，我们展示了为实验放置在场景中的不同物体。请注意，按照定义，所有物体对模型而言都是未见过的，因为用于训练模型的第一人称视角视频来自网络上的现有数据集，并不涉及任何实验室内数据。

5. 结果

本节将讨论无条件模型和目标条件模型的结果，随后进行失败分析。我们展示用于可视化的定性结果，以及在多次试验中的定量评估。

5.1. 无条件生成结果

对于无条件生成结果，给定一个场景，模型会预测一系列动作，并由机器人执行。我们根据机器人执行的轨迹是否对应于场景中人类可能会进行的合理交互来评估模型。例如，关闭的门可以被打开，水果碗可以在桌子上移动；也就是说，给定一个场景，物体状态变化存在一个合理分布。图6展示了机器人在不同初始场景中执行的预测轨迹可视化结果。

2. https://sites.google.com/view/human-0shot-robot

表1：无条件模型的定量评估结果，每种物体进行20次试验。对于每种物体，我们采用不同的方式衡量成功与否。对于门、抽屉和烤面包机，当机器人将相应物体从关闭的初始位置打开，或从打开的初始位置关闭时，即视为成功。对于碗，当碗从初始位置移动到另一位置且未倾倒时，即视为成功。对于移动蔬菜任务，当桌子上至少一半的蔬菜被移动时，即视为成功。

	抽屉	Door	烤面包机	碗 蔬菜		平均值
我们的方法	50%	45%	50%	55%	50%	50%
我们的方法（数据较少）	10%	10%	15%	20%	15%	14%

在表1 中，我们针对场景中的不同物体，对无条件模型生成结果进行定量分析，并为不同物体设定不同的成功标准。对于门、抽屉和烤面包机，若机器人改变了物体的状态（从打开变为关闭，或反之），则视为成功。对于碗，若碗从初始位置移动到其他位置且没有倾倒，则视为成功。从表中可以看出，成功率介于 45% 和 55% 之间，这表明模型能够预测出可在场景中执行合理交互的轨迹。图5中，我们对

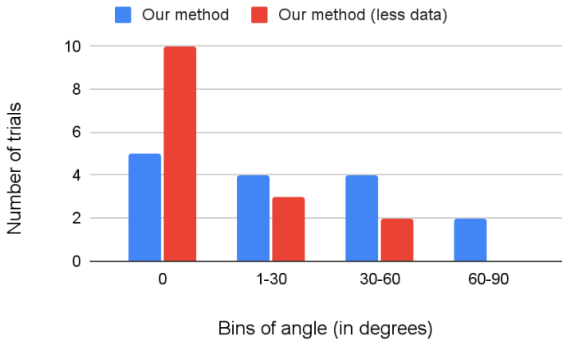

图 5：开门任务的细粒度分析。直方图显示了将门打开到特定角度的试验次数（共 15 次）。

场景中放置一扇关闭的门时的模型预测，并观察到超过 70% 的试验会将门打开到非零角度，这表明物体的状态变化是合理的。

5.2. 目标条件生成结果

除了在给定场景中执行合理的轨迹外，我们还希望了解目标条件模型在生成从初始场景到达指定目标的轨迹方面有多有效。图 7 展示了对应于不同目标图像的轨迹结果。表 2 展示了不同类型目标图像的定量结果。对于每种设置，我们都会随机化物体的初始姿态。例如，对于“打开门”这一目标图像，门的初始位置可能是关闭状态，也可能是半开状态。我们发现，目标条件模型预测的轨迹平均约有37%的成功率；这一比例低于无条件模型，因为到达特定目标比对场景进行无条件改变更困难。与使用较少数据训练的基线方法以及场景流基线方法相比，我们的模型表现更好。在表 3 中，我们针对一个具有堆叠抽屉的场景对模型进行了更细致的分析，发现以目标为条件时，目标条件模型到达目标指定的最终配置的次数多于无条件模型到达相同配置的次数。这表明目标条件化确实会生成与指定目标一致的预测轨迹。

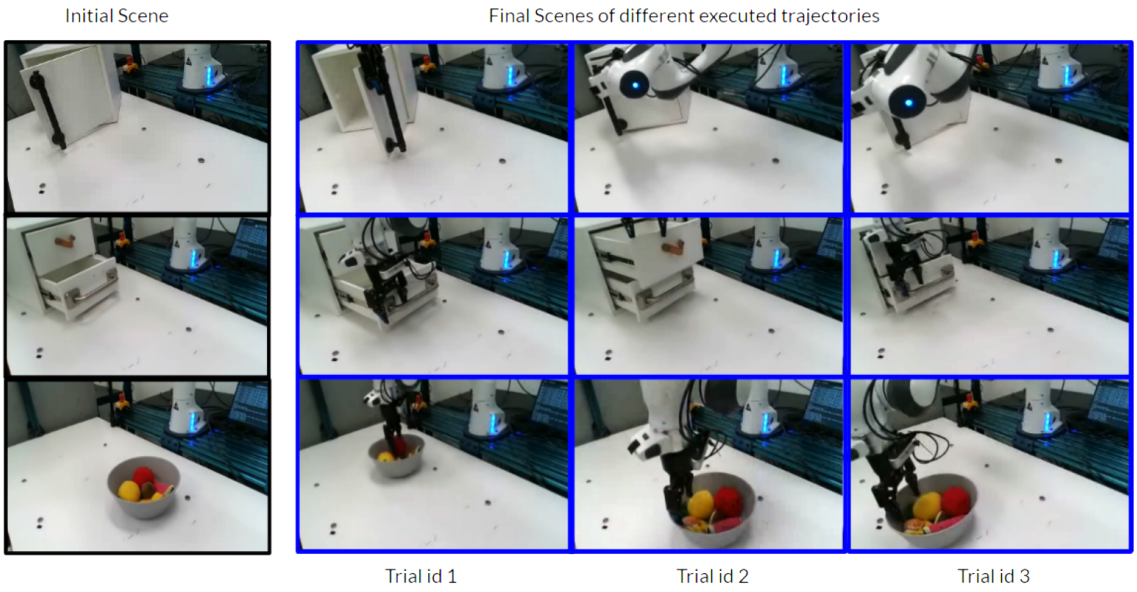

图6：展示无条件模型在场景中执行不同轨迹（试验编号 1、2、3）后对应的最终配置，体现了我们的模型预测出的合理行为的多样性。

表2：目标条件模型的定量评估结果，每个目标类别进行20次试验。每一列表示指定的目标图像类型。在抽屉物体这一项中，我们分别汇总了打开和关闭任务中上层和下层抽屉的结果。

	Open 抽屉	Open Door	关闭 抽屉	关闭 Door	Move Bowl	Move 蔬菜	平均值
我们的方法	35%	30%	35%	40%	45%	35%	37%
我们的方法（较少数据）	10%	5%	10%	15%	15%	10%	11%
三维场景流	15%	15%	5%	10%	15%	10%	12%

表3：给定一个下层抽屉半开、上层抽屉关闭的场景，我们对无条件模型进行10次试验，并统计其中达到某些最终配置的试验次数（共10次）。我们将其与目标条件模型进行比较，后者以对应于某个最终配置的目标作为条件（每个目标进行10次试验）。

已达到最终配置/目标	顶部打开	底部打开	底部关闭
无条件（共 10 次试验）	1/10	3/10	2/10
目标条件（每项 10 次试验）	3/10	4/10	4/10

5.3. 失败分析

对于无条件模型，我们观察到两种主要失败模式：机器人在初始时间步未能与物体接触（约占所有失败的 40%）；以及机器人虽然成功接触物体，但随后的运动不可行，导致机器人卡住（约占所有失败的 60%）。第二种失败模式对应于预测轨迹与相应物体的可供性不匹配，例如试图向外拉门，或按压

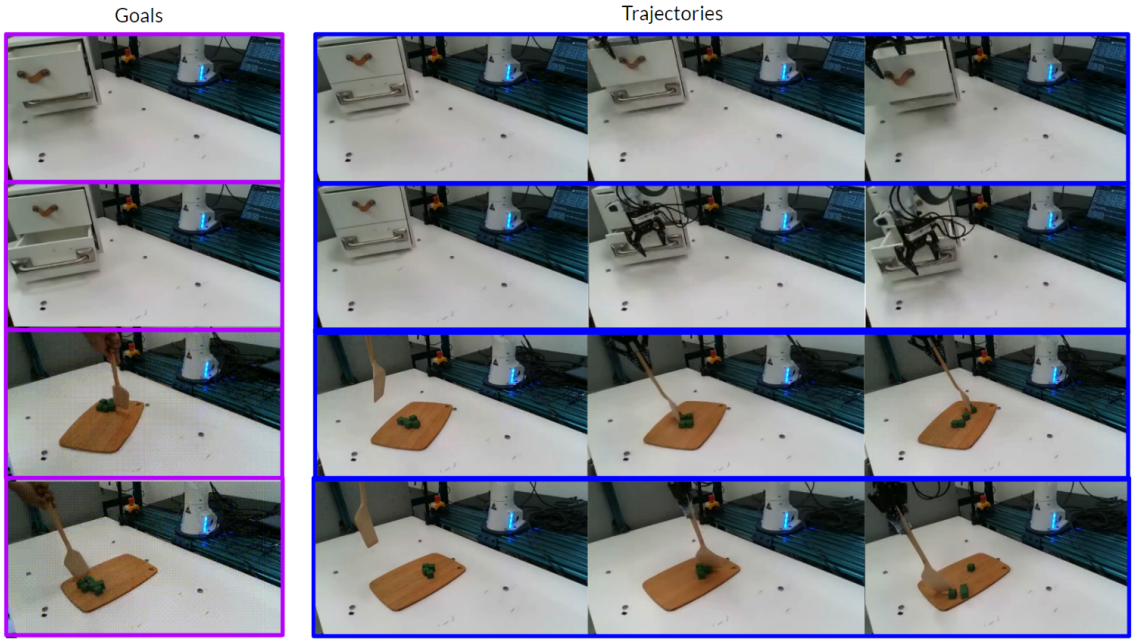

图 7：图中展示了不同的目标条件机器人评估，其目标对应于左侧所示的目标。每一行对应一条子采样轨迹，分别展示了上层和下层抽屉的打开、门的打开与关闭，以及蔬菜在桌子上的移动。

碗的侧面，导致碗在桌子上倾倒。除了上述失败外，对于目标条件模型，我们还观察到第三种失败模式：执行轨迹看似合理（例如打开了较近的抽屉），但与指定目标不符（例如目标显示上层抽屉已打开，而执行轨迹却打开了下层抽屉）。这约占模型失败案例的 40%。

6. 讨论与结论

本文提出了一种方法，可从被动人类视频中预测合理的动作轨迹，从而在给定包含物体的场景时实现零样本机器人交互。我们先从网络上的人类视频中学习一个模型，以预测与智能体无关的动作轨迹；随后将预测结果转换为适用于机器人的具身形态，并在机器人工作空间中直接零样本执行这些运动，无需使用任何实验室内数据进行微调。对于抽屉、烤箱、门和水果碗等日常物体，我们发现预测的轨迹能够实现合理的交互，成功率约为50%；例如，打开关闭着的抽屉，以及移动桌上的碗。我们还进一步开发了该模型的目标条件版本，以场景的初始图像和目标图像为条件；该版本同样使用互联网上的第一人称人类视频进行训练，其中目标图像是视频片段的最终画面。当目标条件模型部署在机器人工作空间中时，我们观察到，从初始场景到达指定目标的成功率约为40%；例如，从门半开的初始场景出发，使门完全打开。我们认为，我们提出的场景条件动作轨迹生成方法有助于理解仅从被动人类视频中提取动作表征的能力边界，以及这些表征在无需微调的情况下对机器人操作的实用性。

参考文献

Pieter Abbeel 和 Andrew Y Ng。通过逆强化学习进行学徒学习。2004年。Yusuf Aytar、Tobias Pfaff、David Budden、Thomas Paine、Ziyu Wang 和 Nando de Freitas。通过观看 YouTube 玩高难度探索游戏。发表于*NeurIPS*，2018年。Seungryul Baek、Kwang In Kim 和 Tae-Kyun Kim。借助神经渲染，推动基于RGB的稠密三维手部姿态估计的发展。发表于*CVPR*，2019年。Shikhar Bahl、Abhinav Gupta 和 Deepak Pathak。真实环境中的人类到机器人的模仿。*arXiv*预印本 *arXiv:2207.09450*，2022年。Michael Bain 和 Claude Sammut。行为克隆框架。发表于《*Machine Intelligence*》，1995年。Gedas Bertasius、Heng Wang 和 Lorenzo Torresani。视频理解是否只需时空注意力？发表于*ICML*，第2卷，第4页，2021年。Mariusz Bojarski、Davide Del Testa、Daniel Dworakowski、Bernhard Firner、Beat Flepp、Prasoon Goyal、Lawrence D Jackel、Mathew Monfort、Urs Muller、Jiakai Zhang 等。端到端自动驾驶汽车学习。*arXiv*，2016年。Adnane Boukhayma、Rodrigo de Bem 和 Philip HS Torr。野外图像中的三维手部形状与姿态。发表于*CVPR*，2019年。Samarth Brahmbhatt、Ankur Handa、James Hays 和 Dieter Fox。Contactgrasp：基于接触的多指功能性抓取合成。*arXiv*，2019年。Berk Calli、Aaron Walsman、Arjun Singh、Siddhartha Srinivasa、Pieter Abbeel 和 Aaron M Dollar。机器人操作研究的基准测试：YCB物体与模型数据集及基准测试协议。*arXiv*，2015年。Matthew Chang、Arjun Gupta 和 Saurabh Gupta。通过观看 YouTube 视频实现语义视觉导航。发表于*NIPS*，2020年。Yu-Wei Chao、Wei Yang、Yu Xiang、Pavlo Molchanov、Ankur Handa、Jonathan Tremblay、Yashraj S Narang、Karl Van Wyk、Umar Iqbal、Stan Birchfield 等。DexYCB：用于捕捉手部抓取物体行为的基准。发表于*CVPR*，2021年。Mark Chen、Alec Radford、Rewon Child、Jeffrey Wu、Heewoo Jun、David Luan 和 Ilya Sutskever。基于像素的生成式预训练。发表于国际机器学习会议，第1691–1703页。PMLR，2020年。Dima Damen、Hazel Doughty、Giovanni Maria Farinella、Sanja Fidler、Antonino Furnari、Evangelos Kazakos、Davide Moltisanti、Jonathan Munro、Toby Perrett、Will Price 等。扩展自我中心视觉：EPIC-KITCHENS数据集。发表于欧洲计算机视觉会议（*ECCV*）论文集，第720–736页，2018年。

Pradipto Das、Chenliang Xu、Richard F Doell 和 Jason J Corso。寥寥数语描述千帧视频：通过潜在主题和稀疏物体拼接对视频进行语言描述。载于*IEEE*计算机视觉与模式识别会议论文集，第2634–2641页，2013年。

Fernando Dela Torre、Jessica Hodgins、Adam Bargteil、Xavier Martin、Justin Macey、Alex Collado 和 Pep Beltran。卡内基梅隆大学多模态活动（CMU-MMAC）数据库指南。2009年。

Laura Downs、Anthony Francis、Nate Koenig、Brandon Kinman、Ryan Hickman、Krista Reymann、Thomas B McHugh 和 Vincent Vanhoucke。Google 扫描物体：一个高质量的三维扫描家居用品数据集。*arXiv*预印本 *arXiv:2204.11918*，2022年。

Chelsea Finn、Tianhe Yu、Tianhao Zhang、Pieter Abbeel 和 Sergey Levine。通过元学习实现单样本视觉模仿学习。载于《机器人学习会议》，第 357–368 页。PMLR，2017年。

Justin Fu、Katie Luo 和 Sergey Levine。通过对抗式逆强化学习学习鲁棒奖励。《arXiv》，2017年。

Antonino Furnari 和 Giovanni Maria Farinella。你会期待什么？利用滚动-展开长短期记忆网络和模态注意力预测自我中心动作。载于《IEEE/CVF 国际计算机视觉会议论文集》，第 6252–6261 页，2019年。

Guillermo Garcia-Hernando、Shanxin Yuan、Seungryul Baek 和 Tae-Kyun Kim。基于 RGB-D 视频和三维手部姿态标注的第一人称手部动作基准。《CVPR》，2018年。

Guillermo Garcia-Hernando、Edward Johns 和 Tae-Kyun Kim。利用估计的手部姿态和残差强化学习实现基于物理的灵巧操作。《arXiv》，2020年。

Liuhao Ge、Zhou Ren、Yuncheng Li、Zehao Xue、Yingying Wang、Jianfei Cai 和 Junsong Yuan。基于单张 RGB 图像的三维手部形状与姿态估计。发表于*CVPR*，2017年。

Mohit Goyal、Sahil Modi、RishabhGoyal 和SaurabhGupta。将人手作为交互式物体理解的探针。发表于*IEEE/CVF*计算机视觉与模式识别会议，第 3293–3303 页，2022年。

Raghav Goyal、Samira Ebrahimi Kahou、Vincent Michalski、Joanna Materzynska、Susanne Westphal、Heuna Kim、Valentin Haenel、Ingo Fruend、Peter Yianilos、Moritz Mueller-Freitag 等。《The" something something"视频数据库：用于学习和评估视觉常识》。发表于 *IEEE* 国际计算机视觉会议，第 5842–5850 页，2017年。

Kristen Grauman、Andrew Westbury、Eugene Byrne、Zachary Chavis、Antonino Furnari、Rohit Girdhar、Jackson Hamburger、Hao Jiang、Miao Liu、Xingyu Liu等。Ego4d：通过3,000小时的自我中心视频环游世界。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第18995–19012页，2022年。

Shreyas Hampali、Mahdi Rad、Markus Oberweger 和 Vincent Lepetit。Honnotate：一种用于手部和物体姿态三维标注的方法。发表于*CVPR*，2020年。

Yana Hasson、Gul Varol、Dimitrios Tzionas、Igor Kalevatykh、Michael J Black、Ivan Laptev 和 Cordelia Schmid。学习手部与被操控物体的联合重建。发表于 *CVPR*，2019年。Yisheng He、Wei Sun、Haibin Huang、Jianran Liu、Haoqiang Fan 和 J. Sun。PVN3D：用于六自由度姿态估计的深度逐点三维关键点投票网络。*CVPR*，2020年。Jonathan Ho 和 Stefano Ermon。生成对抗模仿学习。发表于*NeurIPS*，2016年。Yinlin Hu、Joachim Hugonot、Pascal Fua 和 Mathieu Salzmann。分割驱动的六维物体姿态估计。*CVPR*，2019年。Umar Iqbal、Pavlo Molchanov、Thomas Breuel Juergen Gall 和 Jan Kautz。通过潜在2.5D热图回归进行手部姿态估计。发表于*ECCV*，2018年。Dmitry Kalashnikov、Jacob Varley、Yevgen Chebotar、Benjamin Swanson、Rico Jonschkowski、Chelsea Finn、Sergey Levine 和 Karol Hausman。MT-Opt：大规模连续多任务机器人强化学习。*arXiv*预印本 *arXiv:2104.08212*，2021年。Wadim Kehl、Fabian Manhardt、Federico Tombari、Slobodan Ilic 和 Nassir Navab。SSD-6D：让基于RGB的三维检测和六维姿态估计再度出色。*ICCV*，2017年。Dominik Kulon、Riza Alp Guler、Iasonas Kokkinos、Michael M. Bronstein 和 Stefanos Zafeiriou。野外环境中的弱监督网格卷积手部重建。发表于*CVPR*，2020年。Yin Li、Miao Liu 和 James M Rehg。旁观者眼中的世界：在第一人称视频中联合学习注视与动作。发表于 欧洲计算机视觉会议论文集 *(ECCV)*，第619–635页，2018年。Fangchen Liu、Zhan Ling、Tongzhou Mu 和 Hao Su。基于状态对齐的模仿学习。发表于*ICLR*，2020年。Shaowei Liu、Hanwen Jiang、Jiarui Xu、Sifei Liu 和 Xiaolong Wang。利用时间交互进行半监督三维手部与物体姿态估计。发表于*CVPR*，2021年。Shaowei Liu、Subarna Tripathi、Somdeb Majumdar 和 Xiaolong Wang。从第一人称视角视频预测联合手部运动和交互热点。发表于 *IEEE/CVF*计算机视觉与模式识别会议论文集，第3282–3292页，2022年。Ajay Mandlekar、Yuke Zhu、Animesh Garg、Jonathan Booher、Max Spero、Albert Tung、Julian Gao、John Emmons、Anchit Gupta、Emre Orbay等。Roboturk：通过模仿学习机器人技能的众包平台。发表于机器人学习会议，第879–893页。PMLR，2018年。Kaichun Mo、Leonidas J Guibas、Mustafa Mukadam、Abhinav Gupta 和 Shubham Tulsiani。Where2Act：从像素到可铰接三维物体的动作。发表于 *IEEE/CVF* 国际计算机视觉会议论文集，第6813–6823页，2021年。Tushar Nagarajan、Christoph Feichtenhofer 和 Kristen Grauman。从视频中定位以人为中心的人物-物体交互热点。发表于 *IEEE/CVF* 国际计算机视觉会议论文集，第8688–8697页，2019年。Andrew Y Ng、Stuart J Russell等。逆强化学习算法。2000年。

Deepak Pathak、Parsa Mahmoudieh、Guanghao Luo、Pulkit Agrawal、Dian Chen、Yide Shentu、Eva Shelhamer、Jitendra Malik、Alexei A. Efros 和 Trevor Darrell。零样本视觉模仿。I *ICLR*，2018年。Xue Bin Peng、Angjoo Kanazawa、Jitendra Malik、Pieter Abbeel 和 Sergey Levine。Sfv：强化学习基于视频的身体技能。*TOG*，2018年。Xue Bin Peng、Erwin Coumans、Tingnan Zhang、Tsang-Wei Lee、Jie Tan 和 Sergey Levine 通过模仿动物学习敏捷机器人运动技能。*arXiv*，2020年。Dean A Pomerleau。Alvinn：一种基于神经网络的自主陆地车辆。发表于*NeurIPS*，1989年。Yuzhe Qin、Yueh-Hua Wu、Shaowei Liu、Hanwen Jiang、Ruihan Yang、Yang Fu 和 Xiaolong Wang。Dexmv：从人类视频进行灵巧操作的模仿学习。*arXiv*预印本 *arXiv:2108.05877*，2021年。Mahdi Rad 和 Vincent Lepetit。Bb8：一种可扩展、精准且能抵御部分遮挡的方法，用于在不使用深度信息的情况下预测具有挑战性的物体的三维姿态。*ICCV*，2017年。Yu Rong、Takaaki Shiratori 和 Hanbyul Joo。Frankmocap：快速单目三维手部和身体通过回归和积分进行动作捕捉。*arXiv*预印本 *arXiv:2008.08324*，2020年。St´ephane Ross 和 Drew Bagnell。用于模仿学习的高效约简。发表于*AISTATS*，2010年。Stuart Russell。面向不确定环境的学习智能体。1998年。Karl Schmeckpeper、Annie Xie、Oleh Rybkin、Stephen Tian、Kostas Daniilidis、Sergey Levine，以及Chelsea Finn. 从观测和交互中学习预测模型。*arXiv*，2019年。Karl Schmeckpeper、Oleh Rybkin、Kostas Daniilidis、Sergey Levine 和 Chelsea Finn。强化学习与视频：结合离线观测与交互。*arXiv*，2020年。Pierre Sermanet、Corey Lynch、Yevgen Chebotar、Jasmine Hsu、Eric Jang、Stefan Schaal、Serge Levine 和 Google Brain。时间对比网络：从视频中进行自监督学习。2018年Dandan Shan、Jiaqi Geng、Michelle Shu 和 David Fouhey。理解互联网规模下的接触人手。发表于*CVPR*，2020年。Lin Shao、Toki Migimatsu、Qiang Zhang、Karen Yang 和 Jeannette Bohg。Concept2robot：学习从指令和人类示范中提取操作概念。发表于机器人：科学与系统会议，2020年。Pratyusha Sharma、Lekha Mohan、Lerrel Pinto 和 Abhinav Gupta。多次交互让事情变得更容易(mime)：用于模仿的大规模示范数据。*arXiv*，2018年。Pratyusha Sharma、Deepak Pathak 和 Abhinav Gupta。第三人称视觉模仿学习 vi解耦分层控制器。发表于*NIPS*，2019年。Kenneth Shaw、Shikhar Bahl 和 Deepak Pathak。Videodex：从互联网视频中学习灵巧操作发表于第六届机器人学习年会。Maximilian Sieb、Zhou Xian、Audrey Huang、Oliver Kroemer 和 Katerina Fragkiadaki。Graph结构化视觉模仿。在*CoRL*，2020年。

Aravind Sivakumar、Kenneth Shaw 和 Deepak Pathak。机器人隔空操控：通过观察 YouTube 上的人类来学习机器人手部模仿器。*arXiv*预印本 *arXiv:2202.10448*，2022年。Laura Smith、Nikita Dhawan、Marvin Zhang、Pieter Abbeel 和 Sergey Levine。Avid：通过对人类视频进行像素级转换来学习多阶段任务。*arXiv*，2019年。Shuran Song、Andy Zeng、Johnny Lee 和 Thomas Funkhouser。野外抓取：通过低成本示范学习六自由度闭环抓取。《机器人学与自动化快报》，2020年。Adrian Spurr、Jie Song、Seonwook Park 和 Otmar Hilliges。跨模态深度变分手部姿态估计。在*CVPR*，2018年。Zachary Teed 和 Jia Deng。Raft-3d：利用刚体运动嵌入进行场景流估计。在*IEEE/CVF*计算机视觉与模式识别会议论文集，第8375–8384页，2021年。Faraz Torabi、Garrett Warnell 和 Peter Stone。基于观察的行为克隆。*arXiv*，2018a。Faraz Torabi、Garrett Warnell 和 Peter Stone。基于观察的生成对抗模仿。*arXiv*，2018b。Sundar Vedula、Simon Baker、Peter Rander、Robert Collins 和 Takeo Kanade。三维场景流。在第七届*IEEE*国际计算机视觉会议论文集，第2卷，第722–729页。IEEE，1999年。Limin Wang、Yuanjun Xiong、Zhe Wang、Yu Qiao、Dahua Lin、Xiaoou Tang 和 Luc Van Gool。时序片段网络：迈向深度动作识别的良好实践。在欧洲计算机视觉会议，第20–36页。施普林格，2016年。Yu Xiang、Tanner Schmidt、Venkatraman Narayanan 和 Dieter Fox。PoseCNN：一种用于杂乱场景中六自由度物体姿态估计的卷积神经网络。*arXiv*，2018年。Haoyu Xiong、Quanzhou Li、Yun-Chun Chen、Homanga Bharadhwaj、Samarth Sinha 和 Animesh Garg。通过观察学习：从人类视频中对操作技能进行实体模仿。*arXiv*，2021年。Yufei Ye、Abhinav Gupta 和 Shubham Tulsiani。你手里拿着什么？对手中通用物体进行三维重建。在 *IEEE/CVF*计算机视觉与模式识别会议论文集，第3895–3905页，2022年。Sarah Young、Dhiraj Gandhi、Shubham Tulsiani、Abhinav Gupta、Pieter Abbeel 和 Lerrel Pinto。让视觉模仿变得简单。在机器人学习会议 *(CoRL)*，2020a。Sarah Young、Dhiraj Gandhi、Shubham Tulsiani、Abhinav Gupta、Pieter Abbeel 和 Lerrel Pinto。让视觉模仿变得简单。*arXiv*，2020b。Christian Zimmermann 和 Thomas Brox。学习从单张 RGB 图像估计三维手部姿态。在 *CVPR*，2017年。

7. 附录

7.1. 实验详情

在实验中，我们考虑了图 3 所示的五种不同日常物体，这些物体共有十种不同的语义任务。可能的任务包括：在砧板上移动蔬菜、打开门、关闭门、打开上层抽屉、关闭上层抽屉、打开下层抽屉、关闭下层抽屉、打开烤箱、关闭烤箱，以及将水果碗移过桌面。这些物体和任务涵盖了不同类型的运动，例如工具滑动（砧板）、绕垂直铰链旋转（门）、绕水平铰链旋转（烤箱）以及推动（水果碗）。对于砧板任务，我们将运动限制在桌面平面内，垂直方向的运动限制在距桌面 1 厘米以内（以免锅铲碰到桌面并从夹爪上脱落）。这些不同可能行为的结果旨在展示所提出方法的通用性，即生成无条件行为（任务事先未知时）和目标条件行为（任务由目标图像指定时）。

我们在场景中使用一台 IntelRealsene 相机，输入预测模型的是 RGB 图像观测。在目标条件设置下，目标图像由同一台相机拍摄。相机固定且经过标定。机器人是一台 7自由度 Franka Emika Panda 机械臂，由逆运动学控制器控制。Franka 机械臂底座固定，且其位置已知。末端执行器是一个双指自适应 Robotiq 夹爪。

对于基线方法，我们考虑一种场景流方法 Vedula 等人 (1999)；Teed 和 Deng (2021) 提出了一种使用 RAFT3D 的基线方法Teed 和 Deng (2021) 用于预测初始图像与目标图像之间的场景流场，并利用主导流方向引导机器人的运动。该基线方法使用深度图像（来自同一台相机的 RGBD）计算场景流，因此所需信息比我们的方法更多。为检验在多样化数据上进行训练的重要性，我们将我们的方法与其一个变体进行比较：该变体仅使用全部训练数据的 30% 进行训练，其他训练和评估条件均保持不变。

7.2. 方法的更多细节

模型架构详见第3.1.1节。图 2 展示了无条件模型。目标条件模型与之非常相似，只是额外将一张目标图像作为输入，其特征为 f_g。该图像由 Transformer 编码器编码为 z_g^e。编码中额外加入了一个位置嵌入维度，用于区分初始图像嵌入和目标图像嵌入。其余架构与第3.1.1节所述相同。对于 Transformer，嵌入维度为 512，所有模块的丢弃率均为 0.1。编码器和解码器的 CVAE 网络由两层多层感知器实现。该架构采用自回归方式，即使用先前的动作作为输入来预测当前动作。按照自回归生成的标准做法，在训练期间，我们在每个时间步输入先前的真实动作；在推理期间，则输入先前预测的动作。

使用丰富的互联网视频训练完整体手部位姿预测模型 $p_\psi(a_{1:T}|o_1)$ 后，我们将其直接部署于实验室中的机器人操作任务。机器人通过固定摄像机观察场景图像，并可选择接收一张作为预测模型输入的目标图像。为了使用模型 $a_{1:T}$ 预测的动作来驱动机器人，我们需要将

这些动作转换到机器人的世界坐标系中，并将每个动作 a_t 视为末端执行器目标位姿，再使用低层控制器执行相应的运动。

场景中的相机经过了标定，因此内参矩阵 I 和外参矩阵 $[R, T]$ 已知。世界坐标系位于机器人基座处（机器人基座与桌面处于同一高度），且已知桌面相对于相机的高度，该高度近似恒定。给定相机拍摄的场景 o_1，模型会预测增量动作 $a_{1:T}$，我们将其转换为绝对动作（见第 3.1.1 节），并通过逆投影变换，将动作从相机坐标系变换到机器人的世界坐标系。我们的实验中，预测时域为 $T = 7$。获得动作序列的世界坐标 $\{(X_t,\ Y_t,\ Z_t,\ \alpha_t,\ \beta_t,\ \gamma_t)\}_{t=1}^{T}$ 后，我们使用逆运动学控制器执行相应运动，使末端执行器在每个时间步到达所需的位置和方向。逆运动学控制器的位置误差阈值为 10%，方向误差阈值为 20%，以避免机器人因尝试到达难以实现的预测位姿而卡住。